АНАЛИЗ ТЕКСТА МЕТОДОМ ИЗВЛЕЧЕНИЯ КОЛЛОКАЦИЙ

Джаримбетов.Т.Б

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Хмелева А.В. – канд. тех. наук, доцент

Исследование посвящено анализу эссе студентов БГУИР на тему «Я – программист» с применением метода извлечения коллокаций. Цель – выявить значимые пары слов и интерпретировать их смысл. Анализ 150 текстов выявил сочетания: «востребованная профессия» (t-тест 4,10) и «программный код» (t-тест 4,34), отражающие интерес к карьере и технике. Метод эффективен, но малый объём данных ограничивает выводы. В дальнейшем планируется расширение анализа.

Современные технологии обработки текстов позволяют извлекать из них скрытые закономерности и смысловые связи. Одним из таких подходов является анализ коллокаций. Коллокация – это устойчивое словосочетание, содержащее синтаксические или семантические связи, например, «высокая зарплата», где слова часто встречаются вместе, показывая, что автор связывает эти понятия. При автоматическом выделении коллокаций из текста рассматриваются два подхода: на основе частоты вхождения слов и словосочетаний в рассматриваемый текст (частотный подход) и на основе структуры предложения. Такие сочетания встречаются чаще, чем можно ожидать случайно, и помогают понять, какие темы или идеи важны для автора. Это делает коллокации полезным инструментом в лингвистике и социологии [1].

Настоящее исследование посвящено анализу эссе студентов Белорусского государственного университета информатики и радиоэлектроники (БГУИР) на тему «Я – программист». Цель работы – выявить значимые пары слов вида «существительное + прилагательное» и определить, что они говорят о восприятии студентами своей будущей профессии. Для этого применялись статистические методы: t-тест, PMI и планируется использование критерия χ^2 в будущем. Т-тест проверяет, насколько вероятно, что два слова встречаются вместе не случайно. PMI (Pointwise Mutual Information) измеряет силу связи между словами. χ^2 (критерий хи-квадрат) позволяет оценить значимость сочетаний, сравнивая наблюдаемые и ожидаемые частоты [2].

Для работы были собраны 150 эссе студентов БГУИР, обучающихся на специальности «Программная инженерия». Общий объём текстов составил около 30 000 слов. На первом этапе тексты были подготовлены: удалены заглавные буквы, проведена токенизация с использованием инструмента word_tokenize (часть библиотеки NLTK для анализа текстов), исключены малозначимые слова, такие как «и» или «на», а слова приведены к базовой форме с помощью pymorphy2 (например, «программистов» преобразовано в «программист»). Подобные подходы к обработке текстов с применением программных инструментов описаны в работе Митина и Евдокименко [4]. Затем проводился поиск пар «существительное + прилагательное» с расстоянием от 0 до 2 слов. Например, в фразе «очень востребованная профессия» расстояние между «востребованная» и «профессия» равно 1.

Компьютерный эксперимент на Python включал следующие исследования:

1. Анализ расстояний 0-1-2: подсчитали частоту пар в окне из 3 токенов. Например, «востребованная профессия» – 20 (freq_0 = 15, freq_1 = 3, freq_2 = 2).

2. Частота и пороговый вход: минимальная частота ≥ 5, общее число биграмм N≈15,000.

3. Пороговое значение: t-тест ≥ 1,96 (=0,05):

$$t = \frac{\rho_{\text{биграм}} - \rho_{\text{сущ}} \cdot \rho_{\text{прил}}}{\sqrt{\frac{\rho_{\text{бигр}} \cdot \left(1 - \rho_{\text{бигр}}\right)}{N}}}, \tag{1}$$

где $\rho_{\text{биграм}}$ – вероятность появления пары слов (например, «востребованная профессия»), $\rho_{\text{сущ}}$ – вероятность встретить существительное (например, «профессия»), $\rho_{\text{прил}}$ – вероятность встретить прилагательное (например, «востребованная»), N – общее число пар слов. Если t-тест больше 1,96, пара считается значимой [2].

4. Расчет PMI, чтобы понять, насколько слова связаны:

$$\text{PMI} = \log_2 \left(\frac{\rho_{\text{бигр}}}{\rho_{\text{сущ}} \cdot \rho_{\text{прил}}} \right). \tag{2}$$

Положительное PMI говорит, что слова встречаются вместе чаще, чем случайно [2].

В будущем предполагается использовать χ^2– ещё один способ проверки. Он сравнивает, сколько раз слова встретились вместе, с тем, сколько раз это могло бы быть случайно. Значение χ^2 рассчитывается по выражению:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}, \quad (3)$$

где O_i – реальное число встреч пары в текстах (наблюдаемая частота), E_i – сколько раз пара должна была встретиться случайно (ожидаемая частота), $\sum$ – сумма по всем парам. Чем больше χ^2, тем сильнее связь между словами [2].

Проведенный анализ показал, какие пары слов чаще всего встречаются в эссе студентов. Результаты сведены в таблицу 1.

Таблица 1 – Частота и значимость пар слов в эссе студентов

Пара слов	Частота	Freq_0	Freq_1	Freq_2	t-тест	PMI
Востребованная профессия	20	15	3	2	4.10	3.90
Основа программный	19	19	0	0	4.34	7.81
Сеть программный	14	0	14	0	3.73	8.20
Высокая зарплата	15	12	2	1	3.50	3.20

Частотный анализ выявил, что «востребованная профессия» (частота = 20) и «основа программный» (частота = 19) – лидеры. Технические пары чаще идут без разрыва (freq_0), а прагматические допускают расстояние. Т-тест для всех пар больше 1,96, значит, они не случайны. PMI показывает силу связи: например, у «сеть программный» PMI равно 8,20, что говорит о сильной связи [2].

Дальнейший анализ полученных результатов выявил, что значат эти пары и что их можно поделить на две группы. Первая – про карьеру, например, «востребованная профессия» и «высокая зарплата». Это показывает, что студенты думают о будущем: им важно, чтобы работа была нужной и хорошо оплачивалась. Вторая группа – про технику, например, «основа программный» и «сеть программный». Это говорит, что студентам интересно само программирование. Данный результат совпадает с тем, что пишут другие исследователи, например, в [3].

Метод коллокаций позволил выяснить, о чём думают студенты БГУИР, описывая свою будущую профессию. Пары «востребованная профессия» (t-тест 4,10) и «программный код» (t-тест 4,34) свидетельствуют, что для них важны карьера и техника. Данный способ анализа оказался эффективным, но объём данных составил всего 150 текстов, что недостаточно для масштабных выводов. В дальнейшем планируется увеличить количество текстов и применить χ^2 для повышения точности анализа. Полученные результаты могут помочь БГУИР лучше понять ожидания студентов и улучшить образовательный процесс.

Список использованных источников:

1. Sinclair, J. Corpus, concordance, collocation. Oxford University Press, 1991.

2. Manning, C. D., Schütze, H. Foundations of statistical natural language processing. MIT Press, 1999.

3. Ненаусников, К. В., Кулешов, С. В. Алгоритм автоматического выделения коллокаций из текста. Известия вузов. Приборостроение, 2019, т. 62, № 11, с. 976–981.

4. Митина, О. В., Евдокименко, А. С. Методы анализа текста: методологические основания и программная реализация. Вестник ЮУрГУ, Серия «Психология», 2010, № 40, с. 29–38.